Enrichment of genomic resources and identification of simple sequence repeats from medicinally important *Clausena excavata*

Doo Young Bae¹ · Sang Mi Eum¹ · Sang Woo Lee¹ · Jin-Hyub Paik¹ · Soo-Yong Kim¹ · Mihyun Park¹ · Changyoung Lee¹ · The Bach Tran² · Van Hai Do² · Jae-Yun Heo^{3,4} · Eun-Soo Seong⁵ · Il-Seop Kim⁶ · Ki-Young Choi⁷ · Jin Sung Hong⁸ · Rahul Vasudeo Ramekar⁹ · Sangho Choi¹ · Jong-Kuk Na⁷

Received: 2 August 2017 / Accepted: 6 February 2018 / Published online: 15 February 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

To broaden and delve into the genomic information of *Clausena excavata*, an important medicinal plant in many Asian countries, RNA sequencing (RNA-seq) analysis was performed and a total of 16,638 non-redundant unigenes (≥ 300 bp) with an average length of 755 bp were generated by de novo assembly from 17,580,456 trimmed clear reads. The functional categorization of the identified unigenes by a gene ontology (GO) term resulted in 2305 genes in the cellular component, 5577 in the biological processes, and 8056 in the molecular functions, respectively. The top sub-category in biological processes was the metabolic process with 4374 genes. Among annotated genes, 3006 were mapped to 123 metabolic pathways by the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway analysis tool. The search for simple sequence repeats (SSRs) resulted in 845 SSRs from 749 SSR-containing unigenes and the most abundant SSR motifs was AAG/CTT with 179 occurrences. Twelve SSR markers were tested for cross transferability among five *Clausena* species; eight of them exhibited polymorphism. Taken together, these data provide valuable resources for genomic or genetic studies of *Clausena* species and other relative studies. The transcriptome shotgun assembly data have been deposited at DDBJ/EMBL/GenBank under the accession GGEM000000000.

Keywords RNA-seq · Medicinal plant · *Clausena* species · SSR marker

Introduction

Clausena excavata Burm. f. is a perennial shrub and distributed widely in Southeast Asia, where it is well known as a medicinal plant. It has high antioxidant properties, which are very tightly correlated to the ability to reduce oxidative damage and to inhibit inflammatory conditions by scavenging

Doo Young Bae, Sang Mi Eum, and Sang Woo Lee have contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-018-1162-x>) contains supplementary material, which is available to authorized users.

✉ Jong-Kuk Na
jongkook@kangwon.ac.kr

Sangho Choi
decoy0@kribb.re.kr

¹ International Biological Material Research Center, Korea Research Institute of Bioscience and Biotechnology, Yuseong-gu, Daejeon 34141, Republic of Korea

² IEBR, Vietnam Academy of Science and Technology (VAST), Cay Giay, Ha Noi, Vietnam

³ Agriculture and Life Sciences Research Institute, Kangwon National University, Chuncheon 24341, Republic of Korea

⁴ Present Address: Department of Plant Science, Gangneung-Wonju University, Chuncheon 24341, Republic of Korea

⁵ Department of Medicinal Plants, Suwon Women's University, Suwon 16632, Republic of Korea

⁶ Department of Horticulture, Kangwon National University, Chuncheon 24341, Republic of Korea

⁷ Department of Controlled Agriculture, Kangwon National University, Chuncheon 24341, Republic of Korea

⁸ Division of Bioresource Sciences, Kangwon National University, Chuncheon 24341, Republic of Korea

⁹ Department of Applied Plant Sciences, Kangwon National University, Chuncheon 24341, Republic of Korea

free radicals (Khanna et al. 2007). *C. excavata* has the highest antioxidant activities among the genus *Clausena*, because it contains the highest quantity of phenolic compounds and other pharmacologically active compounds such as furanocoumarins (Ito et al. 2000), flavonoid glycosides (He et al. 2000), and carbazole alkaloid (Sunthitikawinsakul et al. 2003b). The leaves and stems of *C. excavata* have long been used as folk medicine for various disorders such as cough, wounds, fever, headache, and detoxification. Recent studies reported that the extract of *C. excavata* also had various activities including antibacterial (Sunthitikawinsakul et al. 2003a), antifungal (Kumar et al. 2012), antiplatelet (Wu et al. 1994), antitumor (Sharif et al. 2011), and anti-HIV-1 properties (Sunthitikawinsakul et al. 2003a).

Despite the important use of *C. excavata* as folk medicine, its genomic studies have been carried out poorly. Thus, there are very limited sequences available for phylogenetic and population studies. To date, only a handful of DNA and RNA sequences of *C. excavata* can be found in the NCBI database and most of them are DNA barcode markers from phylogenetic and population studies of other plant species, in particular the genus *Citrus* of the Rutaceae family. Available barcode marker genes include chloroplast DNA markers such as maturase K gene, ribosomal protein S4 (Bayer et al. 2009), tRNA-Leu (Bayer et al. 2009; Morton et al. 2003), and ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (Chen et al. 2010). However, there are a few genomic markers from the nucleus. Since *C. excavata* is an important medicinal plant and contains high quantity of phenolic compounds, the addition of more genomic resources can be used to study not only for distinguishing *Clausena* species of the Rutaceae family but also for drug discovery.

With the advent of next-generation sequencing (NGS) technologies, whole genome and RNA-seq transcriptome sequencing can be conducted rapidly and cost-effectively. Thanks to the aid of such new sequencing technologies, genomic studies on non-model organisms that have not been studied in the past have become achievable (Lee et al. 2015). Medicinal plants are not the exception for genomic studies any longer. Recently, the RNA-seq transcriptome analyses of several medicinal plants were reported (Kotwal et al. 2016; Tang et al. 2014; Loke et al. 2016; Rai et al. 2016), revealing

gene structure, gene expression, genomic features, and SSR distribution. Therefore, RNA-seq analysis was performed to broaden genomic resources and to explore the genomic features of *C. excavata*. In addition, SSR distribution and accumulation were examined from *C. excavata* transcriptome data. To our knowledge, this study presents the first whole transcriptome data from the genus *Clausena*, which would be valuable not only for genetic studies but also for understanding the medicinal characteristics and underlying mechanisms of *Clausena* spp.

Materials and methods

Plant materials

To perform the RNA-Seq analysis the fresh leaf tissues of *C. excavata* Burm. f. were collected from Vinh Phuc province, Vietnam during August 2015, and dumped into liquid N₂ immediately after harvest. Samples were carefully transferred into RNAlater solution by following the manufacturer's instructions (Ambion Ins, USA) for long-term storage. To examine the genetic variations among *Clausena* species, five *Clausena* species were used in this study (Table 1). All plant materials were from dried voucher specimens in a herbarium at the International Biological Material Research Center (IBMRC) of the Korea Research Institute of Bioscience and Biotechnology, Daejeon, Republic of Korea. The voucher specimens were deposited from Ghana, Laos, Bangladesh, China, and Vietnam (Table 1).

Preparation of libraries for RNA-Seq

The frozen samples were grounded with a pestle and mortar using liquid N₂ to isolate total RNA using a TRIzol reagent according to the instructions of the manufacturer (ThermoFisher Scientific, Seoul, Korea). The total RNAs were checked using an RNA Pico Chip on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) to estimate the relative purity and concentration. A 10 µg of the total RNA was used for mRNA isolation using oligo-dT beads and sheared randomly for the preparation of the

Table 1 *Clausena* species used in this study

No.	Voucher code ^a	Species	Location (Latitude, Longitude)
1	BL-K 230	<i>C. anisata</i> (Willd.) Hook. f. ex Benth.	Accra, Ghana (5.603717, - 0.186964)
2	LK 697	<i>C. harmandiana</i> (Pierre) Guillaumin	Vientiane province, Laos (18.573918, 102.609936)
3	EDB-KRIBB 401	<i>C. heptaphylla</i> Wight & Arn.	Bandarban, Bangladesh (21.831100, 92.368632)
4	GD-2013-0064	<i>C. lansium</i> (Lour.) Skeels	Guangzhou, China (23.129110, 113.264385)
5	VK 0926	<i>C. excavata</i> Burm. f.	Vinh Phuc, Vietnam (21.360881, 105.547437)

^aIBMRC voucher code ID

Illumina RNA-seq library. The sheared mRNAs were used to generate a cDNA library, followed by the adapter ligation at 3' A overhang for a Illumina Hiseq 2500 analysis. The quality of cDNA library was checked by an Agilent DNA 1000 chip (Agilent Technologies, Santa Rosa, USA), and then sequencing was performed using the Illumina Hiseq 2500 (Illumina, USA). The processing of fluorescent images into sequences, base-calling, and quality value calculations were performed using the Illumina data processing pipeline (version 1.8).

Annotation and gene ontology (GO) analysis

For annotation, the raw reads were filtered by removing the adapter sequences, empty reads, low-quality reads, and reads with less than a Phred quality score of 20 using NGS tool kits and Ion-Torrent server as described previously (Jung et al. 2016). To increase gene identification efficiency, transcriptome de novo assembly was conducted using three assemblers, CLC Genomics Workbench (ver. 3.7.1), Velvet-Oases (ver. 1.1.04-ver. 0.1.21), and Trinity (release 20110519) with various k-mer lengths. A default k-mer value (25-mer) was used for the assembly with CLC (Moreton et al. 2014). For the assembly by Velvet-Oases and Trinity, different k-mer values were applied to obtain the best results from 21 to 79 for Velvet-Oases (Moreton et al. 2014; Schulz et al. 2012) and 25 to 33 for Trinity (Wang and Gribkov 2017). All transcripts by each assembly tool at various k-mer values were merged separately by each assembly tool for further process. As Oases does not cluster assembled contigs, CD-HIT-EST was used to cluster the contigs with an identity more than 90% and coverage of 100% (Li and Godzik 2006). All resulting data sets were merged into a single assembly by collapsing identical or near-identical contigs into single unigenes using CD-HIT-EST. Then, unigenes were annotated by running BLAST with a cut-off *E* value of 10^{-6} against the NCBI non-redundant (NR) protein database (<http://www.ncbi.nlm.nih.gov>). Blast2GO was used to obtain GO annotation of unigenes based on the BLASTX hit against a NR database with a cut-off *E* value of 1.0×10^{-6} (Conesa et al. 2005). The obtained GO terms were classified using WEGO software (Ye et al. 2006).

DNA isolation

Genomic DNA was extracted from silica dried leaves or herbarium specimens using the DNeasy Plant Mini Kit (Qiagen, Seoul, Korea) according to the manufacturer's protocol. DNAs were first evaluated on agarose gel and further confirmed by the PCR amplification of the nuclear ribosomal internal transcribed spacer region using the universal DNA barcode primers, ITS4 (White et al. 1990) and LEU1 (Vargas et al. 1998).

SSR mining and sequence comparison

A total of 16,638 *C. excavata* unigenes generated by de novo assembly from the RNA-seq data were used for SSR mining. SSR detection and primer design were conducted using the Standalone EST Microsatellite Mining and Analysis Tool, SEMAT (Asari et al. 2014). SSRs of mono- to hexa-nucleotide repeats with a minimum of five repeats and 12 bp were identified and analysed. For SSR confirmation, 50 SSRs were randomly selected, and their flanking primers were used for PCR amplification. PCR reaction was performed in a 25 μ L reaction volume containing 50 ng of genomic DNA, 0.2 μ M of each forward and reverse SSR primer (Supplemental Table S1), and 0.25 unit of *Ex-Taq* DNA polymerase (Takara, Seoul, Korea) under the following conditions: an initial denaturation at 95 °C for 3 min, followed by 35 cycles of denaturation at 94 °C for 30 s, annealing at 56 °C for 30 s, polymerization at 72 °C for 30 s, and a final extension at 72 °C for 3 min. The PCR products were verified in 5% agarose gel. SSR primers that produced PCR products of expected size were further used to study SSR transferability among five *Clausena* species.

To compare variations in genic regions of five *Clausena* species, four SSR primers (ctg120, ctg924, and ctg1797, ctg6135; Supplemental Table S1) that produced monomorphic PCR products in all species were selected, and their PCR products were sequenced. Because sequencing of PCR products from ctg6135 primer set failed, obtained sequences from three SSR primer pairs (ctg120, ctg924, and ctg1797) were cleaned, combined into one single sequence, and used for sequence comparison among five *Clausena* species using the Molecular Evolutionary Genetics Analysis (MEGA7) software (Kumar et al. 2016). A dendrogram was reconstructed using the UPGMA method with maximum-composite likelihood and 1000 bootstrap replications using MEGA7 software.

Results

Sequencing and de novo assembly

A total of 20,697,088 raw sequence reads with a length of 2,607,833,088 bp were generated from the paired-end sequencing using the HiSeq 2500 platform (Illumina, FC-401-4003). After removal of the low-quality reads, 17,580,456 trimmed clear reads with a length of 2,143,847,087 bp were used for de novo sequence assembly, which resulted in 16,638 assembled unigenes with an average length of 755 bp. The total length of all unigenes was 12,557,892 bp with 43.7% of GC content, and the longest unigene was 4065 bp. The most abundant unigenes, 3451 (20.74%), were found between 300 and 400 bp range,

Fig. 1 Length distribution of unigenes from the RNA-seq of *C. excavata*

Fig. 2 Top species with high hits of unigenes from the RNA-seq of *C. excavata*

followed by 2415 (14.52%) between 401 and 500 bp, 1923 (11.56%) between 501 and 600 bp, 1604 (9.64%) between 601 and 700 bp, 1268 (7.62%) between 701 and 800 bp, and 1117 (6.71%) between 801 and 900 bp (Fig. 1). More than 50% of the genes belonged between 300 and 700 bp.

Sequence annotation

For annotation, 16,638 assembled unigenes were searched for similarity against the NCBI non-redundant (NR) protein database (<http://www.ncbi.nlm.nih.gov>). Among assembled unigenes, 16,101 genes exhibited a significant similarity to proteins from other organisms, but 537 genes were not similar to other known proteins. Among the annotated genes, 11,680 genes (70%) were similar to proteins from *Citrus sinensis*, 1198 genes (7%) to proteins from *Theobroma cacao*, and 451 (3%) genes to proteins from *Populus euphratica* (Fig. 2).

Functional classification of unigenes by GO analysis

From the GO analysis, a total of 2305 genes were classified into cellular component, 5577 into biological processes,

and 8056 into molecular functions. The majority of genes in the molecular function were classified into two categories (Fig. 3): catalytic activity with 4128 genes (39.5%) and binding with 5209 genes (49.9%). Among the biological processes, the top three categories were the metabolic process with 4374 genes (33.2%), cellular process with 3729 (28.3%), and single-organism process with 2454 (18.6%). The top four categories with the most genes in the cellular component were the cell component with 1445 genes (33.8%), membrane with 1092 (25.5%), organelle with 923 (21.6%), and macromolecular complex with 712 (16.6%), respectively (Fig. 3).

Analysis of genes involved in metabolic pathway

Since *C. excavata* has been well known to contain pharmacologically active compounds including various phenolic compounds, a search for genes involved in the biosynthesis of secondary metabolites was attempted through KEGG metabolic pathway analysis tool. Dozens of genes were found to be involved in biosynthetic pathways or metabolism of terpenoid, ubiquinone, and riboflavin (Table 2). Twenty-three genes were identified to be involved in terpenoid backbone biosynthesis, followed by ten genes in riboflavin metabolism, and three genes in ubiquinone biosynthesis (Table 2). A further in-depth analysis of these genes would assist with the understanding of the biosynthesis of various pharmaceutical compounds in *C. excavata*.

SSR mining and distribution of SSR motif types

The search for perfect SSRs that were longer than 12 bp identified 845 SSRs. The number of SSR-containing unigenes was 749, of which 47 contained more than one SSR. Among the SSRs identified, the most abundant motif types were tri-nucleotide repeats with 659 (78%), followed by di- and mono- with 116 (13.7%) and 53 (6.3%), respectively (Table 3). The most abundant mono-nucleotide repeats were A/T repeats with 51 SSRs, which were significantly higher than C/G repeats with only two occurrences. The AG/CT repeats exhibited the highest occurrence among di-nucleotide repeats, while the most abundant tri-nucleotide repeats were AAG/CTT repeats with 179 occurrences, followed by ATC/ATG repeats with 117 occurrences (Table 4).

Cross-amplification and genetic relationships of five *Clausena* species

The primer 3 that were integrated into SEMAT-32 program generated 551 primer sets from 749 SSR-containing unigenes. The primer sets were 10, 47, 491, and 3 for mono-, di-, tri-, and $\geq$ tetra-nucleotide SSRs, respectively (Table 3). Fifty primer sets were randomly selected from

Fig. 3 Gene ontology term distribution of RNA-seq unigenes. Numbers at the end of each bar denote numbers of unigenes involved in the given process

the di- and tri-nucleotide SSRs (Table 3; Supplementary Table S1) and examined for PCR amplification using the genomic DNA of *C. excavata*. Among them, 37 sets produced PCR products, 21 with PCR products of non-specific or unexpected size, and 16 of an expected size (Supplemental Table S1). Among the 16 primer pairs that produced expected PCR products, 12 were used to amplify the PCR product for five *Clausena* species, *C. anisata*, *C. harmandiana*, *C. heptaphylla*, *C. lansium*, and *C. excavata* (Table 1). All 12 primer pairs gave cross-amplification for three of *Clausena* species, *C. harmandiana*, *C. lansium*, and *C. anisata*, whereas only seven primer sets were successful with amplifying the PCR products for *C. heptaphylla*. Among the 12 SSR primer sets, polymorphism was detected in 8 primer pairs: ctg1411, ctg2564, ctg4770, ctg5884, ctg2725, ctg4029, ctg1961, and ctg906 (Table 5). The polymorphism information content (PIC) values were calculated based on the formula, $PIC = 1 - \sum (p_i)^2$

(Botstein et al. 1980). The PIC value varied from 0.21 for ctg2564 to 0.43 for ctg5884 (Table 5).

To compare a difference in genic regions among five *Clausena* species (Table 1), three different SSR primers, ctg120, ctg924, and ctg1797, were used to amplify genic regions that encode endoplasmic reticulum membrane protein-like, reticulon-like protein B2, and protein IRX15-LIKE-like, respectively (Supplemental Fig. S1). PCR fragments from the three primer pairs were sequenced. Trimmed sequences were combined as one single sequence (Supplemental data) and used for analysis of variations in genic regions of five *Clausena* species and two *Citrus* species, *Citrus clementina* and *Citrus sinensis*. Sequence alignment of genic regions revealed various nucleotide changes among tested *Clausena* species (Supplemental Fig. S1). In dendrogram, genic regions of five *Clausena* species were grouped together, but showed distinct separation from *Citrus* species (Fig. 4). Genic regions of *C. excavata* showed a close

Table 2 Unigenes involved in riboflavin, terpenoid, and ubiquinone metabolic/biosynthetic pathways

Pathway/Seq ID	Enzyme	Pathway ID/EC ID
Riboflavin metabolism		Map00740
VN15080	Reductase	ec:1.1.1.193
VN05882, VN01076, VN14916	Synthase	ec:4.1.99.12
VN02620	Kinase	ec:2.7.1.26
VN12677	Synthetase	ec:2.7.7.2
VN06274	Synthase	ec:2.5.1.9
VN11417	Deaminase	ec:3.5.4.26
VN01076, VN14916	Cyclohydrolase II	ec:3.5.4.25
Terpenoid backbone biosynthesis		Map00900
VN09422, VN06254	Synthase	ec:2.3.3.10
VN15840	Decarboxylase	ec:4.1.1.33
VN04911	Synthase (ferredoxin)	ec:1.17.7.1
VN01667, VN02186, VN16134, VN15322	Synthase	ec:2.2.1.7
VN08981, VN13466	Kinase	ec:2.7.1.36
VN01737	Diphosphate synthase	ec:2.5.1.29
VN11902	Oxidase	ec:1.8.3.5
VN03643	5'-diphospho)-2-C-methyl-D-erythritol kinase	ec:2.7.1.148
VN08438, VN01779	Reductoisomerase	ec:1.1.1.267
VN09401	Diphosphate reductase	ec:1.17.1.2
VN06755	2,4-cyclodiphosphate synthase	ec:4.6.1.12
VN07748	4-phosphate cytidyltransferase	ec:2.7.7.60
VN03879	O-methyltransferase	ec:2.1.1.100
VN16636	Delta-isomerase	ec:5.3.3.2
VN11693, VN12505, VN12802	Reductase (NADPH)	ec:1.1.1.34
Ubiquinone biosynthesis		Map00130
VN04484	Synthase	ec:4.1.3.36
VN05006	Dioxygenase	ec:1.13.11.27
VN09113	Synthase	ec:5.4.4.2

Table 3 Identification of perfect SSRs (≥ 12 bp) and SSR primer sets designed from unigenes of *C. excavata*

	Mono-	Di-	Tri-	> Tri-	Total
SSRs identified ^a	53 (6.3%)	116 (13.7%)	659 (78.0%)	17 (2.0%)	845 (100%)
SSR frequency (kb)	236.9	108.3	19.1	738.7	14.9
SSR primers designed	10 (1.8%)	47 (8.5%)	491 (89.1%)	3 (0.5%)	551 (100%)
Primers tested	0	19	31	0	50

^aCompound SSRs were counted as perfect SSRs

relationship with *C. harmandiana* among tested *Clausena* species (Fig. 4).

Discussion

RNA-seq has become increasingly popular in various fields of medicinal plant transcriptome research. Metabolic pathway analysis of medicinal plants and molecular marker development for molecular breeding has been accelerated through RNA-seq transcriptome analysis (Hao et al. 2012),

though most medicinal plants including *C. excavata* still have limited genomic resources available for in-depth studies. Most parts (e.g., leaf, root, flower, and stem) of *C. excavata* are used for various medicinal purposes (Arbab et al. 2012), and in the present study, whole transcriptome analysis of *C. excavata* was conducted using leaf samples by RNA-seq sequencing technology and bioinformatics tools.

A total of 16,638 unigenes were assembled from ~ 2.14 Gb trimmed RNA-seq sequences. Among them, 16,101 unigenes were annotated by showing significant similarity to other proteins in the NCBI non-redundant (NR)

Table 4 SSR distribution by repeat units

Motifs	No. of repeats											Total	Frequency (kb)
	5	6	7	8	9	10	11	12	13	≥ 14			
A/T	–	–	–	–	–	–	–	11	14	26	51	246.2	
C/G	–	–	–	–	–	–	–	–	1	1	2	6278.9	
AC/GT	–	10	4	4	1	–	–	2	–	–	21	598.0	
AG/CT	–	37	17	10	4	1	–	5	–	6	80	157.0	
AT/AT	–	4	1	3	–	3	–	2	–	2	15	837.2	
AAC/GTT	23	12	9	11	1	1	1	–	2	–	60	209.3	
AAG/CTT	90	41	24	18	–	3	3	–	–	–	179	70.2	
AAT/ATT	12	8	9	5	–	–	–	–	–	–	34	369.3	
ACC/GGT	34	19	5	1	2	–	1	–	–	–	62	202.5	
ACG/CGT	12	6	7	3	–	–	–	–	–	–	28	448.5	
ACT/AGT	4	–	2	–	–	–	–	–	–	–	6	2093.0	
AGC/GCT	59	23	9	4	1	–	–	–	–	–	96	130.8	
AGG/CCT	23	14	5	3	–	–	–	–	–	–	45	279.1	
ATC/GAT	64	28	12	7	4	1	1	–	–	–	117	107.3	
CCG/CGG	22	5	3	1	–	1	–	–	–	–	32	392.4	
≥Tetra	13	2	1	–	–	–	–	–	–	–	17	738.7	
Total	356	209	108	70	13	10	6	21	17	35	845	14.9	

Table 5 Transferability of 12 *C. excavata* SSR loci in four *Clausena* species

No	ID	Motif	No. of Alleles					PIC
			Ca ^b	Cha	Che	Cl	Ce	
1	ctg1411	(AG)6	1	1	0	1	1	0.32
2	ctg2564	(AG)6	3	1	1	1	1	0.21
3	ctg4770	(CT)6	1	1	0	1	1	0.32
4	ctg5884	(GA)6	1	2	0	2	1	0.43
5	ctg120 ^a	(TG)6	1	1	1	1	1	0.00
6	ctg2725	(TCC)5	1	2	1	1	1	0.16
7	ctg6135	(AAG)6	1	1	1	1	1	0.00
8	ctg4029	(CGG)6	1	1	0	1	1	0.32
9	ctg1961	(GAC)6	1	1	0	1	1	0.32
10	ctg906	(GTG)6	2	1	1	1	2	0.24
11	ctg924 ^a	(TCC)6	1	1	1	1	1	0.00
12	ctg1797 ^a	(TCG)6	1	1	1	1	1	0.00
Transferability rate (%) ^c			100	100	41.7	100	–	

^aPCR products were sequenced and used for the comparison of genic regions^bCa: *C. anisata*, Cha: *C. harmandiana*, Che: *C. heptaphylla*, Cl: *C. lansium*, Ce: *C. excavata*^cTransferability rate was calculated based on the presence or absence of PCR products

protein database. Most of the annotated unigenes (11,680) were substantially similar to the proteins of *Citrus sinensis*, which can be explained by the fact that both the genus *Citrus* and the genus *Clausena* belong to the Rutaceae family.

The total number of unigenes identified from *C. excavata* is less than those in *Arabidopsis*, rice, and poplar, in which the number of coding genes was reported as being ~ 26,500, ~ 41,000, and 45,000, respectively (Sterck et al. 2007). The number of identified genes of *C. excavata* ranged from ~ 30 to ~ 77% compared to those found in the

RNA-seq data of crops, *Zea mays*, *Glycine max*, *Capsicum annuum*, and *Helianthus annuus* (Garcia-Ortega and Martinez 2015). In addition, it was lower than the number of genes identified from RNA-seq analysis of other medicinal plants such as *Plantago ovata* and *Isatis indigotica*. In *Plantago ovata*, a total of 46,955 unigenes were identified, of which 28,160 genes were annotated (Kotwal et al. 2016). In *Isatis indigotica*, a total of 33,238 unigenes were assembled, of which 28,184 unigenes were annotated (Tang et al. 2014). It is known that numerous genes express only temporally or

Fig. 4 UPGMA dendrogram using the genic regions from the five *Clausena* and two *Citrus* species. The dendrogram was reconstructed by the UPGMA method based on genic regions amplified from the five *Clausena* species using three SSR primers. Two *Citrus* species were used as the outgroup and their corresponding sequences

(XP_006477474.1: Endoplasmic reticulum membrane protein-like; XP_006476478.1: Reticulon-like protein B2, and XP_006487028.1: Protein IRX15-LIKE-like) were retrieved from the NCBI database. Tree was shown with the percentage of replicate trees with 1000 bootstrap replications

spatially, so the number of expressed genes can range based upon when or what sample was collected and used. In addition, some transcripts would be very difficult to detect due to their low expression level. Recently, it was reported that the ratio of undetected coding genes in RNA-seq analysis ranges from 6% on average to 25% maximum (Garcia-Ortega and Martinez 2015).

In *Citrus*, many SSRs were identified from the EST database (Chen et al. 2006; Luro et al. 2008), BAC end sequence (Ollitrault et al. 2010), and genome sequence (Liu et al. 2013). These *Citrus* SSRs were used to investigate phylogeny or interspecific transferability among *Citrus* species and its close relatives. However, no SSR markers have been reported from the genus *Clausena* to date. In the present study, SSR markers were developed first from *C. excavata*, and 12 SSR markers were used for cross transferability in five *Clausena* species (Table 1). Among them, eight SSR markers exhibited polymorphism (Table 5). In addition, the PCR fragments from the SSR markers, ctg1797, ctg924, and ctg120 (marked by asterisks; Table 5) of five *Clausena* species were sequenced and used to examine variations in genic regions. Genic regions of *C. excavata* and *C. harmandiana* have a close relationship among five *Clausena* species (Fig. 4), which is consistent with the previous observation reported by Samuel et al. (2001). In that study, non-coding sequences of *atpB/rbcL* intergenic spacer region were compared among six *Clausena* species including both *C. excavata* and *C. harmandiana*, in which *C. excavata* and *C. harmandiana* exhibited a close relationship (Samuel et al. 2001).

Medicinal plants contain invaluable molecules with therapeutic potential, and thus, they are an important pool for the identification of new drug candidates (Atanasov et al. 2015). Most *Clausena* species as medicinal

plants contain various therapeutic metabolites. It has been reported that *C. excavata* has the highest medicinal activities in anticancer, antioxidant, antimicrobial, anti-inflammatory, and antimalarial tests among six *Clausena* species including *C. anisata*, *C. lansium*, and *C. harmandiana* (Arbab et al. 2012). Genes involved in the metabolite biosynthesis from the RNA-seq analysis of *C. excavata* were searched and 36 genes involved in riboflavin, terpenoid, and ubiquinone metabolic/biosynthetic pathways were found (Table 2). Among them, 13 genes are involved in the biosynthesis or metabolism of riboflavin and ubiquinone, which have an antioxidant activity (Table 2). However, a limited number of genes that are involved in secondary metabolite biosynthesis were identified in the present study. Therefore, to unravel the details of secondary metabolite biosynthesis pathways, more transcriptome analysis and confirmation of genes involved in such pathways are desired in the future.

Conclusion

In the present study, RNA-seq was conducted to delve into the genomic information of *C. excavata*. A total of 16,638 unigenes were generated by de novo assembly. Among them, 16,101 unigenes (96.8%) were annotated, of which 70% were substantially similar to the proteins of *Citrus sinensis*. SSR mining discovered 845 SSRs from 749 unigenes. Twelve SSR markers were used to investigate the cross transferability among five *Clausena* species, most of which showed successful cross-amplification. Additional genomic and SSR information of *C. excavata* will serve as a valuable genomic resource.

Acknowledgements This work was supported by a grant from the National Foundation for Science and Technology Development of ROK (NRF-2016K1A1A8A01939075) and by 2017 research grant from Kangwon National University.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

References

- Arbab IA, Abdul AB, Aspollah M, Abdullah R, Abdelwahab SI, Ibrahim MY, Ali LZ (2012) A review of traditional uses, phytochemical and pharmacological aspects of selected members of *Clausena* genus (Rutaceae). *J Med Plants Res* 6(38):5107–5118
- Asari N, Ramaswamy M, Subbian E, Palliyarakkal M, Malhotra SK, Karun A (2014) Standalone EST microsatellite mining and analysis tool (SEMAT): for automated EST-SSR analysis in plants. *Tree Genet Genom* 10(6):1755–1757. <https://doi.org/10.1007/s11295-014-0785-2>
- Atanasov AG, Waltenberger B, Pferschy-Wenzig EM, Linder T, Wawrosch C, Uhrin P, Temml V, Wang L, Schwaiger S, Heiss EH, Rollinger JM, Schuster D, Breuss JM, Bochkov V, Mihovilovic MD, Kopp B, Bauer R, Dirsch VM, Stuppner H (2015) Discovery and resupply of pharmacologically active plant-derived natural products: a review. *Biotechnol Adv* 33(8):1582–1614. <https://doi.org/10.1016/j.biotechadv.2015.08.001>
- Bayer RJ, Mabblerley DJ, Morton C, Miller CH, Sharma IK, Pfeil BE, Rich S, Hitchcock R, Sykes S (2009) A molecular phylogeny of the orange subfamily (Rutaceae: Aurantioideae) using nine cpDNA sequences. *Am J Bot* 96(3):668–685. <https://doi.org/10.3732/ajb.0800341>
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32(3):314–331
- Chen C, Zhou P, Choi YA, Huang S, Gmitter FG Jr (2006) Mining and characterizing microsatellites from citrus ESTs. *TAG. TAG Theor Appl Genet Theoretische und angewandte Genetik* 112(7):1248–1257. <https://doi.org/10.1007/s00122-006-0226-1>
- Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5(1):e8613. <https://doi.org/10.1371/journal.pone.0008613>
- Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18):3674–3676. <https://doi.org/10.1093/bioinformatics/bti610>
- Garcia-Ortega LF, Martinez O (2015) How many genes are expressed in a transcriptome? Estimation and results for RNA-Seq. *PLoS One* 10(6):e0130262. <https://doi.org/10.1371/journal.pone.0130262>
- Hao D-C, Chen S-L, Xiao P-G, Liu M (2012) Application of high-throughput sequencing in medicinal plant transcriptome studies. *Drug Dev Res* 73(8):487–498. <https://doi.org/10.1002/ddr.21041>
- He H, Zhu W, Shen Y, Yang X, Zuo G, Ha X (2000) Flavonoid glycosides from *Clausena excavata*. *Acta Botanica Yunnanica* 23(2):256–260
- Ito C, Itoigawa M, Katsuno S, Omura M, Tokuda H, Nishino H, Furu-kawa H (2000) Chemical constituents of *Clausena excavata*: isolation and structure elucidation of novel furanone-coumarins with inhibitory effects for tumor-promotion. *J Nat Prod* 63(9):1218–1224
- Jung H, Yoon BH, Kim WJ, Kim DW, Hurwood DA, Lyons RE, Salin KR, Kim HS, Baek I, Chand V, Mather PB (2016) optimizing hybrid de novo transcriptome assembly and extending genomic resources for giant freshwater prawns (*Macrobrachium rosenbergii*): the identification of genes and markers associated with reproduction. *Int J Mol Sci* 17 (5). <https://doi.org/10.3390/ijms17050690>
- Khanna D, Sethi G, Ahn KS, Pandey MK, Kunnumakkara AB, Sung B, Aggarwal A, Aggarwal BB (2007) Natural products as a gold mine for arthritis treatment. *Curr Opin Pharmacol* 7(3):344–351. <https://doi.org/10.1016/j.coph.2007.03.002>
- Kotwal S, Kaul S, Sharma P, Gupta M, Shankar R, Jain M, Dhar MK (2016) De novo transcriptome analysis of medicinally important plantago ovata using RNA-Seq. *PLoS One* 11(3):e0150273. <https://doi.org/10.1371/journal.pone.0150273>
- Kumar R, Saha A, Saha D (2012) A new antifungal coumarin from *Clausena excavata*. *Fitoterapia* 83(1):230–233. <https://doi.org/10.1016/j.fitote.2011.11.003>
- Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33(7):1870–1874. <https://doi.org/10.1093/molbev/msw054>
- Lee BY, Kim HS, Choi BS, Hwang DS, Choi AY, Han J, Won EJ, Choi IY, Lee SH, Om AS, Park HG, Lee JS (2015) RNA-seq based whole transcriptome analysis of the cyclopoid copepod Paracyplopina nana focusing on xenobiotics metabolism. *Comp Biochem Physiol D Genom Proteom* 15:12–19. <https://doi.org/10.1016/j.cbd.2015.04.002>
- Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>
- Liu SR, Li WY, Long D, Hu CG, Zhang JZ (2013) Development and characterization of genomic and expressed SSRs in citrus by genome-wide analysis. *PLoS One* 8(10):e75149. <https://doi.org/10.1371/journal.pone.0075149>
- Loke KK, Rahnamaie-Tajadod R, Yeoh CC, Goh HH, Mohamed-Hussein ZA, Mohd Noor N, Zainal Z, Ismail I (2016) RNA-seq analysis for secondary metabolite pathway gene discovery in *Polygonum minus*. *Genom Data* 7:12–13. <https://doi.org/10.1016/j.gdata.2015.11.003>
- Luro FL, Costantino G, Terol J, Argout X, Allario T, Wincker P, Talon M, Ollitrault P, Morillon R (2008) Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina* Hort ex Tan) to other Citrus species and their effectiveness for genetic mapping. *BMC Genom* 9:287. <https://doi.org/10.1186/1471-2164-9-287>
- Moreton J, Dunham SP, Emes RD (2014) A consensus approach to vertebrate de novo transcriptome assembly from RNA-seq data: assembly of the duck (*Anas platyrhynchos*) transcriptome. *Front in Genet* 5:190. <https://doi.org/10.3389/fgene.2014.00190>
- Morton CM, Grant M, Blackmore S (2003) Phylogenetic relationships of the Aurantioideae inferred from chloroplast DNA sequence data. *Am J Bot* 90(10):1463–1469. <https://doi.org/10.3732/ajb.90.10.1463>
- Ollitrault F, Terol J, Pina JA, Navarro L, Talon M, Ollitrault P (2010) Development of SSR markers from Citrus clementina (Rutaceae) BAC end sequences and interspecific transferability in Citrus. *Am J Bot* 97(11):e124–129. <https://doi.org/10.3732/ajb.1000280>
- Rai A, Yamazaki M, Takahashi H, Nakamura M, Kojoma M, Suzuki H, Saito K (2016) RNA-seq transcriptome analysis of *Panax japonicus*, and its comparison with other Panax species to identify potential genes involved in the saponins biosynthesis. *Front Plant Sci* 7:481. <https://doi.org/10.3389/fpls.2016.00481>

- Samuel R, Ehrendorfer F, Chase MW, Greger H (2001) Phylogenetic analyses of Aurantioideae (Rutaceae) based on non-coding plastid DNA sequences and phytochemical features. *Plant Biol* 3(1):77–87. <https://doi.org/10.1055/s-2001-11747>
- Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28(8):1086–1092. <https://doi.org/10.1093/bioinformatics/bts094>
- Sharif N, Mustahil N, Mohd Noor N, Sukari M, Rahmani M, Taufiq-Yap Y, Ee G (2011) Cytotoxic constituents of *Clausena excavata*. *Afr J Biotech* 10(72):16337–16341
- Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y (2007) How many genes are there in plants (... why are they there)? *Curr Opin Plant Biol* 10(2):199–203
- Sunthitikawinsakul A, Kongkathip N, Kongkathip B, Phonnakhu S, Daly JW, Spande TF, Nimit Y, Napaswat C, Kasisit J, Yoosook C (2003a) Anti-HIV-1 limonoid: first isolation from *Clausena excavata*. *Phytother Res PTR* 17(9):1101–1103. <https://doi.org/10.1002/ptr.1381>
- Sunthitikawinsakul A, Kongkathip N, Kongkathip B, Phonnakhu S, Daly JW, Spande TF, Nimit Y, Rochanaruangrai S (2003b) Coumarins and carbazoles from *Clausena excavata* exhibited antimycobacterial and antifungal activities. *Planta Med* 69(2):155–157. <https://doi.org/10.1055/s-2003-37716>
- Tang X, Xiao Y, Lv T, Wang F, Zhu Q, Zheng T, Yang J (2014) High-throughput sequencing and De Novo assembly of the *Isatis indigotica* transcriptome. *PLoS One* 9(9):e102963. <https://doi.org/10.1371/journal.pone.0102963>
- Vargas P, Baldwin BG, Constance L (1998) Nuclear ribosomal DNA evidence for a western North American origin of Hawaiian and South American species of *Sanicula* (Apiaceae). *Proc Natl Acad Sci USA* 95(1):235–240
- Wang S, Gribkov M (2017) Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics* 33(3):327–333. <https://doi.org/10.1093/bioinformatics/btw625>
- White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis N, Gelfand D, Sninsky J, White T (eds) *PCR protocols: a guide to methods and applications*. Academic Press Inc., New York, pp 315–322
- Wu CC, Ko FN, Wu TS, Teng CM (1994) Antiplatelet effects of clausine-D isolated from *Clausena excavata*. *Biochem Biophys Acta* 1201(1):1–6
- Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, Wang J (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* 34((Web Server issue)):W293–W297. <https://doi.org/10.1093/nar/gkl031>